Coding for Demographic Categories in the Creation of Legacy Corpora: Asian American Ethnic Identities

Coding for demographic categories in the creation of legacy corpora: Asian American ethnic identities

Abstract:

A set of shared coding conventions for speaker ethnicity is necessary for open-source data sharing and cross-study compatibility between linguistic corpora. However, ethnicity, like many other aspects of speaker identity, is continually negotiated and reproduced in discourse, and therefore a challenge to code representatively. This paper discusses some of the challenges facing researchers who want to use, create, or contribute to existing corpora that are annotated for the ethnic identity of a speaker. We specifically problematize the macro-social label 'Asian American' and propose that researchers should consider different levels and types of specificity of 'Asianness' in order to ensure that the corpora best represent the reality of ethnic identity in the community sampled. This is particularly important given the limited incorporation of different Asian groups in most existing linguistic research (cf. Reyes and Lo 2009). We argue that more rigorous coding for Asian American ethnicities in corpora will improve the utility of archived corpora and enhance sociolinguistic research on language variation and ethnic identity.

1. Introduction

A set of shared coding conventions for speaker ethnicity[1] is necessary for open-source data sharing and cross-study compatibility between linguistic corpora. However, ethnicity, like many other aspects of speaker identity, is continually negotiated and reproduced in discourse, and therefore a challenge to code in a way that truly represents the options of ethnic identification that are important to a community and its members. This paper joins others in this issue in discussing some of the challenges facing researchers who want to use, create, or contribute to existing corpora that are annotated for the ethnic identity of a speaker. We specifically problematize the macro-social label given to American's fastest-growing racial group, 'Asian American', and propose strategies for ensuring that corpora which include Asian Americans best represent the

[1] In this paper we are not making an explicit distinction between 'ethnicity' and 'race'. For more information on this complex issue, see Omi and Winant (1994) and Fought (2006).

reality of a speaker's ethnic identity. This is particularly important given the insufficient incorporation of different Asian groups in sociolinguistic studies in the United States[i] (cf. Reyes and Lo 2009). For example, as far as we are aware, there are only a relative handful of quantitative sociophonetic studies that explicitly examine the use of regional phonological/phonetic features by American-born individuals of Asian heritage, and most of these consider only East Asian heritage groups (Mendoza-Denton and Iwai 1993; Lee 2000; Wong 2007; Hall-Lew 2009; Becker 2010; Ito 2010; Wong 2010; Kaiser 2011; Newman and Wu 2011; Wong 2012; Hall-Lew 2013; Wong and Hall-Lew 2014). Related work in linguistic anthropology has contributed to the breadth of heritage ethnicities represented in sociolinguistics more broadly, with a particular focus on communities of South and Southeast Asian Americans (e.g., Reyes 2005, 2007; Shankar 2008; Bucholtz 2009 [2004]; Shankar 2011). The 2010 U.S. Census finds that the population of Asian-identified Americans is growing at a faster rate than the total U.S. population (U.S. Census Bureau 2012b). Incorporating more rigorous coding of Asian American ethnicities in North American corpora can help linguistics keep pace with social change, as well as improving the utility of archived corpora and enhancing research on language and ethnic identity.

This paper reviews three examples: one shows a context in which the broad label 'Asian' may actually reflect local identity well. In another, individuals' varying degrees of orientation towards an ethnic label may account for a significant amount of linguistic variation. A third example further complicates the picture by showing how ethnic orientation can shift within a single interview. Even in the case where 'Asian' is the preferred community label, in practice that label is actively negotiated, often framed as including certain groups and excluding others. In some instances, individuals may use 'Asian' to refer restrictively to East Asians, excluding South and Southeast Asians explicitly. In other instances, individuals may use 'Asian' as a pan-ethnic term encompassing a wide range of identities. (See Section 2 below for some of these possible identities). As a result, coding for ethnic identity may also be a dynamic process for the researcher, in that some key intra-group differentiations or the precise meaning of 'Asian' may only become apparent during fieldwork or after fieldwork has been completed.

This paper discusses techniques for collecting and representing nuanced information about ethnicity, including the use of questionnaires to represent identity on a quantifiable scale. We propose that each speaker within legacy corpora could be tagged, given appropriate supporting evidence, with multiple codes that are relevant to ethnicity and will offer examples from our own research. We focus on Asian Americans for exemplification, but the suggestions may have broader applicability. More rigorous coding for ethnic identity across all corpora will allow for more comparable work as well as increased data sharing across linguistics.

2. Coding for Asian Ethnic Identities in Metadata for North American corpora

In *Sociolinguistics and Corpus Linguistics*, Baker observes, "[c]lassifications based on concepts like ethnicity, social class, sexuality and sex can be problematic, resulting in over-simplifications, stereotyping or reinforcing prejudice" (2010: 9). But rather than abandoning corpus linguistics, Baker argues that socially aware quantitative methods can exist in tandem with constructionist approaches. (See also Biber's (2012) review of Baker (2010) for a comparison of corpus linguistics and sociolinguistics.) The present paper is a test case for this perspective – specifically, how socially representative can corpus metadata be?

Every sociolinguist building a corpus will confront challenges with respect to coding speakers for ethnic identity. One inherent complication has to do with the fact that ethnicity is both a macro-level social process and micro-level individual practices (see, for instance, Omi and Winant (1994)). The question for researchers then is how to represent both the macro- and micro-level information in the metadata. Reference to U.S. Census categories is one common approach to representing the macro-label. Given that census categories are often evoked during naturalistic discursive interactions by individuals themselves, they often provide a very reasonable and useful option for coding. However, the changing nature of those census categories presents a number of complexities.

For example, the terms of reference for Americans of Asian ancestry have differed for nearly every U.S. Census. The first category applying to people of Asian ancestry was 'Chinese', used to describe three individuals in 1830 (Shinagawa and Kim 2008), and

used alone throughout the mid-1800s until the additional category of 'Japanese' was added (in 1870 in California and 1890 in other states; U.S. Census Bureau 2012a). Over the 20th century, Census categories were added which continued to differentiate Asian Americans with respect to their ancestral nation of origin ('Filipino' in 1920, 'Vietnamese' in 1980, etc.). While the broad category of 'Asian' has never been a single Census option (other than 'Other Asian'), it is widely used in written reports generated based on Census data (e.g., U.S. Census Bureau 2012a, b). The current U.S. Census definition of 'Asian' is:

> A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam. It includes 'Asian Indian,' 'Chinese', 'Filipino', 'Korean', 'Japanese', 'Vietnamese', and 'Other Asian'. (U.S. Census Bureau n.d.)

However, just as the difference between 'Europe' and 'Asia' is itself the product of "arbitrary historical boundaries" (Lineback and Gritzner 2013), so too are the groups included in the 'Asian' category changing and likely to change in the future. For example, South Asians (initially, 'Hindu', a term used in the 1920-1940 Census reports) were for a time (1950-1990) classified as 'White'. Additionally, data on Asian groups has long been and is still often presented along with data on 'Native Hawaiian and Other Pacific Islander[s]' (henceforth NHPI; U.S. Census Bureau 2011), although NHPI was officially separated from the category of 'Asian' in 1997 (Washington State Office of Financial Management 2001). The general U.S. Census categories also do not explicitly deal with non-national ethnic groups, such as 'Chinese Vietnamese' or 'Hmong', which maybe salient identities in local communities (Kaiser 2011).

Of increasing importance to Asian American identities, and American ethnic identities more generally, is the difference between what U.S. Census refers to as the "race alone population" and the "race in combination population" (U.S. Census Bureau 2011: 5). This latter designation became available in 1997, and consequently in the 2000 U.S. Census, when a change was made from instructing respondents to "Mark one race

only" to "Mark one or more if necessary" (Washington State Office of Financial Management 2001). In 2010, 2.9% of the total U.S. population chose more than one racial classification on their census form (up from 2.6% in 2000; U.S. Census Bureau 2005). Comparing across "major race group" (U.S. Census Bureau 2011: 9), the group with the highest percentage of people identifying as multiracial was NHPI (55.9%), followed by American Indians and Alaska Natives (henceforth AIAN, 43.8%), and then the Asian group (15.3%). In terms of absolute numbers, most multiracial Americans chose White as their major race group (7.5 million), followed by Black (3.1 million), and then Asian (2.6 million). In other words, in contrast to other major ethnic groups, Asian Americans represent both a relatively high proportion as well as a relatively high number of individuals who identify with more than one ethnicity.

The increasing number of Americans identifying with more than one racial group implicates how speakers' ethnic identity is coded for. Fought (2006: 72) notes that multiracial individuals are commonly found as participants in sociolinguistic studies in the U.S., but that, with some exception (Bucholtz 1995; Gordon 2000), there is a dearth of research on language and multiracial identity.[ii] Sociolinguistic studies and naturalistic corpora from the U.S. are increasingly facing the practical challenge of annotating ethnicity for purposes of statistical analysis and metadata composition. One of the three case studies in this paper will consider an example of a multiracial Asian American, or *Hapa* (a term originally from Hawai'ian used throughout Asian American communities to refer to individuals of partial Asian ancestry). Our suggestion will be that the practical way forward is for researchers to code every identity a speaker makes relevant in the observed interaction, and that 'Mixed' or 'Hapa' are themselves likely candidates for those identities.

3. Case Studies

What follows are three examples from our research that raise the question of how to code for the ethnic identities of Americans of Asian heritage.

3a. When local labels match census labels

Our first example comes from work on a sociophonetic study of a residential neighborhood in San Francisco, California (Hall-Lew 2009, 2013; Wong and Hall-Lew 2014). Recordings with residents of that neighborhood, the Sunset District, comprise a corpus currently under construction, the Sunset Corpus (Hall-Lew 2014). In considering how ethnic identity is entered in the corpus metadata, this example shows that, despite the well-founded criticism that broad labels erase important group-internal diversity, there are cases where those same labels are given local interpretation, warranting or even necessitating the adoption of those labels in corpus annotation.

Relying *a priori* on broad labels to represent ethnic identity runs the risk of masking variation between individuals that may be important for analyses, and also potentially misrepresents the speaker's own sense of self. Mendoza-Denton, in describing the speakers in her work (1997, 1999, 2008), notes that her speakers "who would be classified under one census category—the monolithic 'Hispanic' category—come from many different countries, social-class backgrounds, and prior educations experiences" (1999: 277). Her analyses convincingly show how these sub-group differences correlate with patterns of language use. However, this does not mean that broad categories can never form parts of the ethnographically emergent categories. Census categories are themselves discursive objects, arising either as labels imposed by out-group members or by labels embraced by in-group members. An individual filling out a census form interacts with those labels, choosing among them (or offering their own). In this sense, census categories do represent a particular set of real-world social processes and can be locally relevant. The present example considers one context where the very broad category, 'Asian', is the ethnic label most often used by local residents, and thus should be adopted in corpus annotation. Crucially, the decision to do so should be driven by ethnographic observations, and how the label is interpreted locally should be annotated.

"The term 'Asian American' was first used in the San Francisco Bay Area to spearhead a national movement uniting people of diverse Asian heritages around common causes and interests" (Hall-Lew 2010: 464). As a result, this category, even broader than ancestry-based U.S. Census categories (e.g., 'Chinese'), has local historicity and relevance. Today, the single term 'Asian' is frequent in local discourse, 'American' being dropped for various reasons (either because American identity is presumed, or

because American identity is inaccurate as the referent might not have been born in the U.S. or have U.S. citizenship). As one Sunset business owner said in 2008, "My customer base now is at least fifty percent Asian." This quote represents typical statements that show that 'Asian' is the locally preferred term. Equally important, however, is what that term means. In the Sunset District, 'Asian' might be pan-ethnic, but just as often it refers specifically to those of Chinese heritage (who comprise the majority of residents identifying as Asian, both locally and nationally). The terms may be used synonymously, such as when a local journalist says, "The Sunset is fifty percent Chinese, Asian…" or "…from an Asian American community, Chinese American community standpoint…" or, when speaking about the past, "Asians, Chinese, were probably more in the minority." The equivalence of 'Asian' and 'Chinese' is most apparent when 'Asian' and not 'Chinese' appears in a list of local ethnic contrasts, such as when another resident says "…you're having more Korean churches, Vietnamese churches, Asian churches," with even list intonation suggesting that 'Asian' is not meant as an umbrella term but as a parallel to 'Korean' and 'Vietnamese'. Ongoing linguistic analysis of this community further suggests that Chinese Americans define the range of local phonetic variability, with variation among non-Chinese Asian Americans and Whites falling within that range (Hall-Lew 2009, 2013).

This complex alignment between ethnic identity labels has implications for two distinct levels of corpus metadata: that which characterizes the community that the speakers belong to, however that is defined, and that which characterizes the individual speakers themselves. The description of the community in the Sunset Corpus is two-fold: 1.) that it is predominantly 'Asian', and 2.) that 'Asian' is a term that can either have a pan-ethnic meaning or be used as a synonym for 'Chinese'. Furthermore, the range of 'pan-ethnic' meaning is also variable with respect to non-Chinese ethnicities: it is possibly more likely to include Japanese than Korean, and possibly more likely to include Korean than Filipino, depending on the speaker. While some of this information may be apparent in the discursive data available in the corpus itself (audio files, transcripts), some might only have emerged through ethnographic work, in which case it is particularly vital to include in the metadata. Lastly, to provide maximal description for unforeseen future study, the percentages of each 'ethnic' group (as defined by the U.S.

Census results closest to the time of data collection) comprising the neighborhood population are also included, with ethnographically informed annotations, where appropriate. This description of the neighborhood then implicates how speaker-specific ethnic labels are applied and interpreted. In the Sunset Corpus, each speaker is labeled with as many labels as is deemed necessary to represent their ethnic identity. This will be discussed further in section 3c, below.

In this example, we advocate retaining broad category descriptors for both communities and individuals, as long as the linguist also includes the motivations and justifications for the use of such labels. In short, before employing census categories, it is best to consider their ethnographic validity.

3b. When census labels may be inadequate

Our second example comes from work on American-born Cantonese in New York City by the second author. Wong focused on examining the variation in the use of regional features of New York City English among Cantonese New Yorkers who are often simply coded as members of the same minority group. In particular, she shifted her attention from treating 'Cantonese American' as a fixed and homogeneous ethnic group to looking at how differences in speaker's ethnocultural alignment might be connected to variation in the use of regional linguistic features. Previous social scientific research on Chinese Americans have shown that individuals' ethnic identity and orientation vary greatly in terms of immigration generation, age, socioeconomic status, the ethnic density of their residential community, their social network, and participation in cultural activities and language use (see Tsai et al. 2000; Ying et al. 2008).

In an earlier study, Wong (2007, 2010) examined four New York-born and raised Chinese American females of Cantonese descent, between the ages of 18-29 at the time of data collection in 2006. She measured these four individuals' cultural/lifestyle orientation and their social network (Milroy 1980, 2001) using a set of questionnaires (Tsai et al. 2000; Kirke 2005). Based on the questionnaire responses, she classified the four speakers according to whether they exhibited a balanced/bidimensional orientation toward Chinese and American cultures or a biased/unidimensional orientation towards one culture (Tsai et al. 2000) and whether they exhibited a Chinese-dominant social

network or not (see Wong 2007). Of the four females, two had a non-Chinese dominant social network and a more unidimensional orientation towards an American lifestyle. These are also the speakers who pronounced a raised-BOUGHT vowel (in words like *thought* and *caught*), the variant that is stereotypically linked to the New Yorker persona (Wong 2010; Becker 2011; Wong and Hall-Lew 2014), more frequently. This study indicates that variation in social network and cultural orientation corresponds to differences in the use of regional phonological features.

A more recent study on raised-BOUGHT with a larger sample of New York born Cantonese Americans (N=32) shows similar effects of speakers' cultural orientation (Wong in preparation). The sample consisted of speakers from a wider age range (11-69) and equal numbers of males and females. Information concerning a speaker's orientation towards their heritage culture was gathered through ethnographic interviews and supplemented by ethnographic observations. Speakers were coded for their heritage-cultural orientation (strong vs. moderate vs. weak) along other social and linguistic predictors (such as phonological environments, speaker's age, gender, social class, etc.). Statistic analyses show that variation in the height of BOUGHT was predicted by speaker age. Cantonese New Yorkers who pronounced BOUGHT with raising were all older (born on or before 1985). No younger speakers produced raised-BOUGHT (see also Wong 2012; Wong and Hall-Lew 2014 for similar age effects). However, not all older speakers produced raised-BOUGHT. The height of BOUGHT among older speakers was significantly predicted by speakers' orientation towards their heritage culture. Speakers with weak and moderate orientation towards their heritage culture produced BOUGHT that is more raised than speakers with strong orientation towards their heritage culture.

These findings show that despite often being labeled (or labeling themselves) using the same broad ethnic label ('Chinese American' or 'Asian American'), Cantonese New Yorkers differ in their ethnocultural orientation and alignment, with linguistic consequences. Those Cantonese New Yorkers who desired to dissociate themselves from the immigrant culture and to align themselves more with mainstream culture used a salient regional feature of English in New York more frequently. The use of regional phonological features by members of the same ethnicity to carve out intraethnic and local social distinctions is also seen in other minority groups (e.g., Fought 1999). Meaningful

intra-ethnic variation is revealed only if the researcher moves beyond coding a speaker's ethnic identity with broad census labels only and codes also for those behavioral and attitudinal domains that may impact a speaker's ethnic orientation. Speakers' frequency of use of a heritage language, the ethnic composition of their social networks, and their participation in various aspects of ethnic culture are a few domains commonly measured in social scientific studies of ethnic identity (Keefe 1992; Phinney 1992; Roberts et al. 1999; Tsai et al. 2000, inter alia; Hoffman and Walker 2010; Nagy et al. 2014).

3c. Accounting for complex and fluid identities

The final example is again from San Francisco, California, this time focusing on the complexities that arise in representing an individual's ethnic identity when that individual is actively orients to a multiracial identity and also variably orients to one or more components of that identity. The question is how to best code such a speaker's ethnic identity in the metadata of the Sunset Corpus.

Mickey (a pseudonym) began his 2008 interview by describing his parents' ethnicities: "Mom was White, Dad was Filipino Chinese… Dad…is of mixed background." Despite the fact that his mother's ethnicity was mentioned first, discursively it is clear that Mickey is by his own definition 'Asian'. Speaking of his childhood, he says: "With Asians sometimes they'll drop you off with relatives to be raised village style. And that's how I was done." In the course of a sociolinguistic interview he consistently uses this term in combination with first person pronouns: "Asians, that's the way we are…," "…a lot of Asians, we are conservative anyway," and about his interest in Kung Fu, "[be]cause I'm Asian, I wanted to do something Asian." Mickey thus consistently situates his multiracial heritage as an identity included in the Sunset District discourse of Asian pan-ethnic identity. This is particularly apparent in his orientation to the interviewer (Hall-Lew), who is also multiracial: "…moving up in the world, just like you and I, Asians, moving up in the world."

Yet that alignment with 'Asian' does not negate his clear orientation to different aspects of that Asian identity. He characterizes his childhood as being first more influenced by one ethnicity, then by another: "because he was born there [the Philippines], he [Dad] knew a lot of … Filipino relatives. So as a youngster, five or so, I

knew most of that side. Then I grew into my Chinese side." Reacting to this, the interviewer asked, "How did that happen?" to which Mickey replied, "In the 50s there weren't many Filipinos here, there were more Chinese than there were Filipinos. And Asians tended to stick together in the fifties and sixties." Whereas he earlier presented his father's Filipino-Chinese ethnicity as multiracial, here he emphasized that both ethnicities are Asian. He also remarked on the prevalence of multiracial identities within Asian groups, saying that "a lot of Filipinos say they're half Chinese," immediately following this by framing his identity strictly in genealogical terms, "I'm a quarter." At the same time, however, the Chinese dominance of the 'Asian' identity that obtains for the wider Sunset community appears in another part of Mickey's interview: "…that's what made our Chinese community." To the extent that his identity is variably framed as explicitly Chinese, he states that language learning and linguistic ability are part of that: "I go to China once a year (for) the last ten years … I speak some Cantonese, enough to get by."

Mickey's ethnic identity is not easily reduced to a single descriptor. Despite his dominant self-description as 'Asian', and despite his active participation in the wider 'Asian' community, he orients to various other identities over the course of the interview. For the sake of completeness, and with the knowledge that the scope of future research questions cannot be known at the time of corpus compilation, we suggest here that the best strategy is to include all potentially relevant ethnic descriptors as they arose in the interview (and, where relevant, ethnography). In other words every speaker's metadata entry in the Sunset Corpus has the potential for having more than one label for ethnic identity. While some speakers in the community will just have one (e.g., 'White', with no further ethnic descriptors), others will have at minimum two (e.g., 'Asian' and 'Chinese', or 'White' and 'Irish'). Additional annotations may also be provided for each of these labels. Mickey's entry includes five columns representing ethnic identity: Asian, Mixed,[iii] Chinese, Filipino, and White. This last column requires additional annotation; although his mother was of Irish descent, Mickey never claims this label and actively rejects the label 'White' in the context of the interview. Labels that might describe a speaker's ethnic identity at some level but which are not actively invoked by that speaker are marked with an asterisk in the metadata file (see Figure 1). Note that these labels would be applied prior to any further ethnic orientation analysis that might also be conducted; inclusion of

questionnaire-based data such as those described in 3b will then result in additional annotation regarding each of those ethnic labels. Crucially, the choice of which descriptor(s) to refer to at the analysis stage is a choice that depends on the research questions and objectives, not the construction of the corpus.

Figure 1: Metadata for the speaker set in the Sunset Corpus (Hall-Lew 2014)

Insert Figure 1 around here

The three examples above highlight some of the complexities in annotating ethnic identity information in a corpus. While these examples are meant to demonstrate variation among Asian American communities, in fact they barely scratch the surface: all three come from communities that are primarily Chinese American, and they only come from two cities. Despite their limitations, these cases demonstrate some of the many issues that should be considered by any sociolinguists who plan to work with Asian Americans: (1) that communities that are predominantly oriented to one heritage ethnicity (Chinese) may, for various reasons, label that orientation as pan-ethnic (Asian), (2) that individual members of an etically homogenous ethnic group will personally orient and align more or less to the cultural practices associated with that group, and (3) that all individuals in those communities will orient more or less to their other multiple social identities, and that sometimes those identities will also be ethnic ones. In the next section, we will review some methods for a corpus compiler to collect and represent information on ethnic identity and orientation during their ongoing research or from previously completed fieldwork and how as corpus metadata.

4. Finding and Representing Information on Ethnic Identity and Orientation

Surveys are seemingly straightforward ways to gather information on ethnic identity on some quantifiable scale (Wong 2007; Hoffman and Walker 2010). However, as Nagy et al. (2014) note, there is not a single survey of ethnic identity and orientation that is widely used by sociolinguists. Rather, studies that have used surveys to collect

information among speakers of Asian descent in North America (including both the US and Canada) have adopted questionnaires from psychology and sociology.

The study by Wong (2007, 2010) discussed above used two questionnaires to collect information on ethnic orientation. They are essentially two versions of the same 27 statements that differ only in the reference culture. These statements cover various domains of cultural practices, such as language use and proficiency, media and food consumption, affiliation with cultural groups, and attitudes towards some of these practices. They were selected and adapted from the General Ethnicity Questionnaire developed by Tsai, a social psychologist, and colleagues (Tsai et al. 2000; Tsai n.d.). Speakers were asked to rate, on a three-point scale, how accurate they found each of the statements to be in describing their own lifestyle practices. The points given by a speaker for each of the 27 statements in a questionnaire were then summed up as a general score for that speaker's orientation towards the reference culture. Each speaker received two general scores, one for their orientation towards the Chinese heritage culture, and one for their orientation towards the American culture. These scores are useful for corpus annotation, as speaker-specific metadata, and for analysis, as predictor variables. The scores are also comparable to one another, such as in Wong's (2007, 2010) third score of cultural orientation for each speaker, a score of difference, which was calculated by subtracting an individual's Chinese score from their American score.

Studies in Canada that include Asian Canadians (Hoffman and Walker 2010; Nagy 2013; Nagy et al. 2014) have also employed survey methods. These studies used an ethnic orientation questionnaire to gather information on speaker ethnic identification, language use, attitudes about the importance of ethnic culture, etc. Rather than asking speakers to rate the accuracy/applicability of a set of statements, the questionnaire contains a series of open-ended questions. In this sense, the questionnaire could be easily integrated into the sociolinguistic interview (Labov 1981) as a separate module. Hoffman and Walker (2010), for instance, administered the questionnaire at the end of each interview. A speaker's responses allow for the researchers to code separately for each of the several behavioral and attitudinal aspects of ethnic identity (e.g. heritage language use, ethnicity of networks, etc), which can then be used in corpus annotation or quantitative analysis. Speakers' responses to each question can and should be included in

the metadata of future sociolinguistic corpora. These responses can later function as inputs to a single composite measurement of ethnic orientation, which can be further added to the speaker-specific metadata.

In cases where survey administration is not feasible, such as in the creation of legacy corpora based on previously completed fieldwork, information on ethnic orientation can be gleaned from discursive evidence from the very recordings that comprise the corpus being archives (and from any accompanying ethnographic notes). This includes speakers' bald statements of ethnic self-identification and more context-dependent statements about affiliation with ethnic groups, past and current, and participation in ethnic practices as well as other meta-commentary of self-identification, including narratives of mismatch between the speaker's own sense of self and the way they are received in interaction with others (as seen in the case studies above). In particularly complex cases, like the third case above, a fully representative set of coding may require close discursive analysis of the emergence of ethnic identification through narratives of the past and present (see also Bucholtz 1995). Of course, sociolinguistic corpora will differ between those that contain questions specifically asking about ethnicity and the speaker's ethnic identity, and those that do not have such questions. In the former case, information appropriate for metadata coding can be easily extracted from the content of the interview, whereas in the latter case it cannot.

The general strategy we propose for accounting for incorporating the fluidity of ethnic identity in metadata management is *maximal coding*, the inclusion of all likely and possible labels of ethnic identity for any given speaker. In practice this means that there is no limit on the number of codes representing any single speaker's ethnic identity, and that it is also important to describe the ethnic composition of the speaker's residential community. The challenge, then, is to decide on which labels should be included. Ethnicity, like all aspects of an social identity, is complex for all individuals because it is negotiated in interaction, contingent both on descriptors held by the individual and descriptors encountered by the people that individual encounters and interacts with. The greater the number of differences between an individual's self definition and the (set of) externally imposed definition(s), the greater the complexity in coding. Sociolinguistic research from the former perspective focuses on the connection between ethnographically

derived ethnic labels and language use, whereas work from the latter perspective correlates language use with ethnic labels taken from more abstract sociological surveys, like the U.S. Census. However, data management must remain agnostic to which of these perspectives is more important to sociolinguistic analysis, and therefore to ensure long-term utility of any corpus we suggest that it is necessary to include all potentially relevant ethnic labels, both self- and other-ascribed, into the corpus metadata.

5. Conclusion

North American populations are becoming increasingly diverse with respect to ethnicity, and a rise in the number and diversity of Asians and Asian Americans is part and parcel of this process. On the other hand, there has been a shift within variationist sociolinguistics toward a constructivist view of ethnicity not as objectively definable categories but as sets of cultural practices (Wenger 1998; Brubaker 2002; Ashmore et al. 2004; Brubaker et al. 2004; Bucholtz and Hall 2005). Recognizing that perspective, we have outlined some of the practical challenges and offered some strategies for approaching them. For example, it is vital to always code speaker identity with more detail than just 'Asian' or 'Asian American' (as even the U.S. Census does), and it is simultaneously important to justify any use of broad labels, to recognize that speakers will differ in their orientation to those labels, and to recognize that one speaker's orientation may shift over the course of a single interaction. While the strategy we present for maximal coding gets part way to representing the complexity and hybridity of ethnicity identity, it nonetheless still under-represents its fluidity. The larger question for corpus-based sociolinguistics is therefore more a question of how the field can meet its analytic goals of bridging models of language use with social theory, in this case how to balance static representations of an individual or a community (i.e., corpus metadata) while considering the complex and dynamic reality of our contemporary social world. Although our discussions have focused mainly on Asian American groups, our proposals are equally applicable to the coding of other ethnicities.

Given the rapidly changing demographics of American society, one of our goals in sociolinguistics should be to build corpora that incorporate ethnic groups that have traditionally been overlooked in large-scale dialectological work. At the moment,

relatively little is known about ‘Asian Americans’ with respect to dialect studies, and so it is not entirely clear what levels of ethnicity and ethnic orientation will end up being most important to future sociolinguistic descriptions. While some work in variationist sociolinguistics in North America has confronted the challenge of how to best code for ethnic identity for the purpose of quantitative analysis (see, for example, Fought 2006; Wong 2007; Hoffman and Walker 2010; and articles within Hall-Lew and Yaeger-Dror 2014), it is equally important to consider the implications for data storage, management, and archiving (see also Kendall 2008). The challenge for both linguistic analysis and corpus construction is to recognize the ways in which every speaker may have multiple, variable ethnic identities which are intersecting and potentially negotiated throughout a given recording. Luckily, the different long-term goals between analysis and data management mean that there is more freedom in the latter to provide copious amounts of both fine-grained and coarse-grained metadata. Maximal coding for ethnicity is one strategy for ensuring both the accuracy and longevity of linguistic corpora.

Works Cited

Ashmore, Richard D., Kay Deaux and McLaughlin-Volpe. 2004. An organizing framework for collective identity: Articulation and significance of multidimensionality. Psychological Bulletin 130. 80-114.

Baker, Paul. 2010. Sociolinguistics and Corpus Linguistics. Edinburgh: Edinburgh University Press.

Becker, Kara. 2010. Regional Dialect Features on the Lower East Side of New York City: Sociophonetics, Ethnicity, and Identity. Ph.D. dissertation. New York: New York University.

Becker, Kara. 2011. The social meaning(s) of raised-BOUGHT in New York City: A perceptual approach. Paper presented Paper presented at the the 40th Annual Conference on New Ways of Analyzing Variation (NWAV 40), October 27-30, Washington, DC.

Biber, Douglas. 2012. Review article: Sociolinguistics and Corpus Linguistics (Paul Baker) and Corpus and Sociolinguistics: Investigating Age and Gender in Female Talk (Bróna Murphy). Journal of Sociolinguistics 16. 301-07.

Brubaker, Rogers. 2002. Ethnicity without groups. European Journal of Sociology 43. 163-89.

Brubaker, Rogers, Mara Loveman and Peter Stamatov. 2004. Ethnicity as cognition. Theory and Society 33. 31-64.

Bucholtz, Mary. 1995. From Mulatta to Mestiza: Language and the reshaping of ethnic identity. Gender Articulated: Language and the Socially Constructed Self, ed. by K. Hall and M. Bucholtz, 351-74. New York: Routledge.

Bucholtz, Mary. 2009 [2004]. Styles and stereotypes: Laotian American girls' linguistic negotiation of identity. Beyond Yellow English: Toward a Linguistic Anthropology of Asian Pacific America, ed. by A. Reyes and A. Lo, 21-42. Oxford: Oxford University Press.

Bucholtz, Mary and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. Discourse Studies 7. 585-614.

Dubois, Sylvie and Megan Melançon. 1997. Cajun is dead – Long live Cajun: Shifting from a linguistic to a cultural community. Journal of Sociolinguistics 1. 63-93.

Fought, Carmen. 1999. A majority sound change in a minority community: /u/-fronting in Chicano English. Journal of Sociolinguistics 3. 5-23.

Fought, Carmen. 2006. Language and Ethnicity. Cambridge, UK ; New York: Cambridge University Press.

Gordon, Matthew. 2000. Phonological correlates of ethnic identity: Evidence of divergence? American Speech 75. 115-36.

Hall-Lew, Lauren. 2009. Ethnicity and Phonetic Variation in a San Francisco Neighborhood. Ph.D. dissertation. Palo Alto: Stanford University.

Hall-Lew, Lauren. 2010. Ethnicity and sociolinguistic variation in San Francisco. Language and Linguistics Compass 4. 458-72.

Hall-Lew, Lauren. 2013. 'Flip-flip' and mergers-in-progress. English Language and Linguistics 17. 359-90.

Hall-Lew, Lauren. 2014. The Sunset Corpus, in progress. Online: http://talkbank.org/resources/metamaker/form/metadata/0metadata-thesunse01.

Hall-Lew, Lauren and Malcah (Eds.) Yaeger-Dror. 2014. New perspectives on the linguistic variation and ethnic identity in North America [Special issue]. Language and Communication 35.

Hoffman, Michol and James A. Walker. 2010. Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. Language Variation and Change 22. 37-67.

Ito, Rika. 2010. Accommodation to the local majority norm by Hmong Americans in the Twin cities, Minnesota. American Speech 85. 141-62.

Kaiser, Eden A. 2011. Sociophonetics of Hmong American English in Minnesota. Ph.D. dissertation. Minneapolis: University of Minnesota.

Keefe, Susan Emley. 1992. Ethnic identity: the domain of perceptions of and attachment to ethnic groups and cultures. Human Organization 51. 35-43.

Kendall, Tyler. 2008. On the history and future of sociolinguistic data. Language and Linguistics Compass 2. 332-51.

Kirke, Karen D. 2005. When there's more than one norm-enforcement mechanism: Accommodation and shift among Irish immigrants to New York City. University

of Pennsylvania Working Papers in Linguistics: Selected Papers from NWAV 33 11. 1-12.

Labov, William. 1981. Field methods of the project on linguistic change and variation. Sociolinguistic Working Paper. Number 81. Southwest Educational Development Lab. Online: http://eric.ed.gov/?id=ED250938. Retrieved.

Lee, Hikyoung. 2000. Korean Americans as Speakers of English: The Acquisition of General and Regional Features. Ph.D. dissertation. Philadelphia: University of Pennsylvania.

Lineback, Neal and Mandy Lineback Gritzner. 2013. The georgraphical divisons of Europe and Asia. Geography in the NewsTM. Online: http://newswatch.nationalgeographic.com/2013/07/09/geography-in-the-news-eurasias-boundaries/. Retrieved August 27, 2013.

Mendoza-Denton, Norma. 1997. Chicana/Mexicana Identity and Linguistic Variation: An Ethnographic and Sociolinguistic Study of Gang Affiliation in an Urban High School. Ph.D. dissertation. Palo Alto: Stanford University.

Mendoza-Denton, Norma. 1999. Turn-initial *no*: Collaborative opposition among Latina adolescents. Reinventing Identities: The Gendered Self in Discourse, ed. by M. Bucholtz, A. C. Liang and L. A. Sutton, 273-92. New York: Oxford University Press.

Mendoza-Denton, Norma. 2008. Homegirls: Language and Cultural Practice among Latina Youth Gangs. Oxford: Blackwell.

Mendoza-Denton, Norma and Melissa Iwai. 1993. "They speak more Caucasian": Generational differences in the speech of Japanese-Americans. Paper presented to the Proceedings of the First Annual Symposium about Language and Society-Austin (SALSA), Austin, 1993.

Milroy, Lesley. 1980. Language and Social Networks. Oxford: Basil Blackwell.

Milroy, Lesley. 2001. Social networks. The Handbook of Variation and Change, ed. by J. K. Chambers, P. Trudgill and N. Schilling-Estes, 549-72. Oxford: Blackwell.

Nagy, Naomi. 2013. Heritage Language Variation and Change. Online: http://projects.chass.utoronto.ca/ngn/HLVC/0_0_home.php. Retrieved August 1, 2013.

Nagy, Naomi, Joanna Chociej and Michol Hoffman. 2014. Analyzing ethnic orientation in the quantitative sociolinguistic paradigm. Language and Communication 35. 9-26.

Nagy, Naomi, Nina Aghdasi, Derek Denis and Alexandra Motut. 2011. Null subjects in heritage languages: contact effects in a cross-linguistic context. University of Pennsylvania Working Papers in Linguistics: Selected Papers from NWAV 39 17. 135-44.

Newman, Michael and Angela Wu. 2011. "Do You Sound Asian When You Speak English?": Racial Identification and Voice in Chinese and Korean American's English. American Speech 86. 152-78.

Omi, Michael and Howard Winant. 1994. Racial Formation in the United States: From the 1960s to the 1990s New York: Routledge.

Phinney, Jean S. 1992. The multigroup ethnic identity measure: a new scale for use with diverse groups. Journal of Adolescent Research 7. 156-76.

Reyes, Angela. 2005. Appropriation of African American slang by Asian American youth. Journal of Sociolinguistics 9. 509-32.

Reyes, Angela. 2007. Language, Identity, and Stereotype among Southeast Asian American Youth: The Other Asian. Mahwah, NJ: Lawrence Erlbaum.

Reyes, Angela and Adrienne Lo (eds) 2009. Beyond Yellow English: Toward a Linguistic Anthropology of Asian Pacific America. New York: Oxford University Press.

Roberts, Robert E., Jean S. Phinney, Louise C. Masse, Y. Richard Chen, Catherine R. Roberts and Romero Andrea. 1999. The structure of ethnic identity of young adolescents from diverse ethnocultural groups. Journal or Early Adolescence 19. 301-22.

Shankar, Shalini. 2008. Speaking like a model minority: 'FOB' styles, gender, and racial meanings among Desi teens in Silicon Valley. Journal of Linguistic Anthropology 18. 268-89.

Shankar, Shalini. 2011. Style and language use among youth of the new immigration: Formations of race, ethnicity, gender, and class in everyday practice. Identities: Global Studies in Culture and Power 18. 646-71.

Shinagawa, Larry Hajime and Dae Young Kim. 2008. A Portrait of Chinese Americans. Organization of Chinese Americans and Asian American Studies Program, University of Maryland. Online. Retrieved.

Tsai, Jeanne L. n.d. General Ethnicity Questionnaire. Online: http://psych.stanford.edu/~tsailab/GEQ.htm. Retrieved July 22, 2013.

Tsai, Jeanne L., Yu-Wen Ying and Peter A. Lee. 2000. The meaning of 'being Chinese' and 'being American': Variation among Chinese American young adults. Journal of Cross-Cultural Psychology 31. 302-32.

U.S. Census Bureau. 2005. We the People of More Than One Race in the United States. Census 2000 Special Reports.

U.S. Census Bureau. 2011. 2010 Census Data Results for the Asian Population and the Native Hawaiian and Other Pacific Islander Population. Presentation to the White House Initiative on Asian Americans and Pacific Islanders, May 19, 2011.

U.S. Census Bureau. 2012a. The Asian Population: 2010. 2010 Census Briefs.

U.S. Census Bureau. 2012b. 2010 Census Shows Asians are Fastest-Growing Race Group. [News release]. Online: http://www.census.gov/newsroom/releases/archives/2010_census/cb12-cn22.html. Retrieved August 15, 2013.

U.S. Census Bureau. n.d. Race. Online: http://quickfacts.census.gov/qfd/meta/long_RHI505210.htm. Retrieved August 15,, 2013

Washington State Office of Financial Management. 2001. Understanding census 2000: Race category changes & comparisons. Population Estimates & Projections. Research Brief No. 12.

Wenger, Etienne. 1998. Communities of Practice: Learning, Meaning, and Identity. Cambridge: Cambridge University Press.

Wong, Amy Wing-mei. 2007. Two vernacular features in the English of four American-born Chinese. University of Pennsylvania Working Papers in Linguistics: Selected Papers from NWAV 35 13. 217-30.

Wong, Amy Wing-mei. 2010. New York City English and Second Generation Chinese Americans. English Today 26. 3-11.

Wong, Amy Wing-mei. 2012. The lowering of raised-THOUGHT and the low-back distinction in New York City: Evidence from Chinese Americans. University of Pennsylvania Working Papers in Linguistics: Selected Papers from NWAV 40 18. 157-66.

Wong, Amy Wing-mei. in preparation. Diverse Linguistic Resources and Multidimensional Identities: A Study of the Linguistic and Identity Repertoires of Second Generation Chinese Americans in New York City [Working Title]. Ph.D. dissertation. New York: New York University.

Wong, Amy Wing-mei and Lauren Hall-Lew. 2014. Regional variability and ethnic identity: Chinese Americans in New York City and San Francisco. Language and Communication 35. 27-42.

Ying, Yu-Wen, Meekyung Han and Sandra L. Wong. 2008. Cultural orientation in Asian American adolescents: Variation by age and ethnic density. Youth and Society 39. 507-23.

Figure 1: Metadata for the speaker set in the Sunset Corpus (Hall-Lew 2014)

LLC_Figure1.xls

Sheets | Charts | SmartArt Graphics | WordArt

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Pseudonym**	**YOB**	**Age @ IV**	**MF**	**Ethnicity (A)**	**Ethnicity (B)**	**Ethnicity (C)**	**Ethnicity (D)**	**Ethnicity(E)**	**Immigrant Generation**	**Father born**	**Mother born**
2	April	1989	18	F	White	Slovenian	–	–	–	3	San Francisco	San Francisco
3	Carrie	1967	40	F	Asian	Chinese	Singaporean	Surinamese	–	2	Hong Kong	Singapore
4	Cheri	1942	65	F	White	Irish	–	–	–	3	San Francisco	San Francisco
5	Cindy	1966	41	F	Asian	Chinese	Singaporean	Surinamese	–	2	Hong Kong	Singapore
6	Danny	1963	44	M	White	Polish	–	–	–	2	Slovenia	*unknown*
7	Emiko	1942	65	F	Asian	Japanese	–	–	–	3	San Francisco	San Francisco
8	Emily	1970	37	F	Asian	Chinese	Cantonese	–	–	2	China	China
9	Enid	1932	76	F	Asian	Chinese	Cantonese	–	–	2	Fresno, CA	Hawaii
10	Irene	1983	24	F	Asian	Chinese	Taiwanese	–	–	2	Taiwan	Taiwan
11	Jenny	1949	60	F	Asian	Chinese	Cantonese	–	–	2	China	China
12	John	1977	30	M	Asian	Chinese	Hawaiian	–	–	2	Hawaii	Hong Kong
13	Mary	1978	29	F	White	Irish	–	–	–	5	Sunset	East Coast
14	Maya	1983	24	F	Asian	Mixed	Chinese	Filipino	–	3	Sunset	Sunset
15	Mickey	1944	63	M	Asian	Mixed	Chinese	Filipino	Irish*	2	Phillipinnes	*unknown*
16	Molly	1971	36	F	Asian	Chinese	Northern Mainla	–	–	3	Fresno, CA	Hong Kong
17	Monica	1991	16	F	Asian	Chinese	Cantonese	–	–	2	Shanghai	Canton
18	Pete	1984	23	M	Asian	Chinese	Cantonese	–	–	2	China	Hong Kong
19	Richard	1967	40	M	White	Irish	English	–	–	4	Belfast, Ireland	San Francisco
20	Sal	1962	45	M	Asian	Chinese	Cantonese	–	–	2	China	Hong Kong
21	Sam	1990	17	M	White	Mixed*	Irish	Korean	–	2	*unknown*	Sunset
22	Skylar	1991	16	M	Asian	Chinese	Cantonese	–	–	2	Hong Kong	E. China
23	Vicky	1985	22	F	Asian	Chinese	Shanghainese	–	–	2	Shanghai	Shanghai

Endnotes

[i] There is reason to think that most if not all of the issues considered here hold for Canadian corpora as well (see Hoffman and Walker 2010; Nagy et al. 2011). However, the meaning and relevance of these particular distinctions and categories will necessarily be different in contexts outside of North America. For example, the ethnonym 'Asian' itself connotes an entirely different heritage group in the United Kingdom (South Asians) than it does in North America (East Asians).

[ii] This is aside from studies of those communities defined by distinctly local multiracial identities, e.g. Creole African Americans in Southern Louisiana (e.g., Dubois and Melançon 1997) .

[iii] 'Mixed' is the term that he uses; this might otherwise be 'multiracial' or 'biracial' or 'Hapa', or all of the above, depending on what terms the speaker offers and to what extent the fieldworker can determine the difference between terms intended as synonyms and terms intended as contrasts.